The Semantic Planetary Data System

J. Steven Hughes[1], Daniel J. Crichton[1], Sean Kelly[1], and Chris Mattmann[1]

[1]Jet Propulsion Laboratory
4800 Oak Grove Drive
Pasadena, CA 91109 USA
{steve.hughes, dan.crichton, sean.kelly, chris.mattmann}@jpl.nasa.gov

Abstract. The Planetary Data System (PDS) data model was developed in the late 1980's and models the entities and relationships of interest within the Planetary Science Community. It was developed to both prescribe the metadata to be collected for the planetary science data archive and to design the PDS Catalog, a high level inventory of the data holdings in the archive. This catalog, an inventory of over one thousand data sets and related entities, represents a planetary science ontology. Since the advent of the Web, most of the information on the Web has been designed for human consumption using HTML and the http protocol. Semantic Web languages now allow computer processing and reasoning of web information by computer software. This paper will describe the use of these languages and other semantic web technologies to build the Semantic PDS prototype, an application that allows facet- and text-based search of PDS Catalog information.

1 Introduction

The Planetary Data System (PDS) data model was developed in the late 1980's to model the various entities and relationships of interest within the Planetary Science Community. It was developed to both prescribe the metadata to be collected for the planetary science data archive and to design the data set catalog, a high level inventory of the data holdings in the archive. The data model, implemented in a relational schema for the catalog database, supports sophisticated constraint-based searches for data sets based on their relationships to other modeled entities such as the spacecraft instruments and target bodies that were involved in the collection of the data.

Since the advent of the Web, most of the information on the Web has been designed for human consumption using web technologies such as HTML and the http protocol. The Semantic Web now provides technologies to allow information to be easily read and consumed by computer software. These new technologies such as the XML language, the Resource Description Framework (RDF), and RDF Schema (RDFS) support computer processing and reasoning of web information. This capability however is dependent on the existence of domain ontologies.

The Semantic PDS prototype demonstrates the use of semantic web technologies to capture, document, and manage the PDS data model and to provide intuitive facet- and text-based search for data holdings in the archive. The prototype makes use of the PDS Catalog, an inventory of over one thousand data sets and related entities. The underlying data model was ingested into an ontology tool and then exported as a Resource Description Framework (RDF) Schema file in XML format (RDFS/XML). Data records from the PDS Catalog were then written to an RDF/XML file that conformed to the RDFS/XML specifications. Finally the two files were imported into a web-based semantic search engine that provides facet- and text-based search of data sets, instruments, missions, and other modeled entities. The two files, representing a knowledge base, can also be made available to "semantically aware" software, allowing computers to process and reason about the information.

2 The Planetary Science Ontology

The Planetary Data System (PDS) is the official science data archive for NASA's planetary science community. As such, it contains tens of terabytes of data collected from over thirty years of solar system exploration and will grow exponentially in the next few years. The PDS developed a data model, illustrated in Figure 1 that guides the capture of the information necessary to describe the data and ensure that the data remain scientifically useful for future scientists. Collected and validated using the data model, this information and the science data is submitted to peer review, archived, and distributed to the planetary science community. The data model was also used to design the data set catalog, a high level inventory of the data holdings in the archive. The data model, implemented in a relational schema for the catalog database, supports sophisticated constraint-based searches for data sets based on their relationships to other modeled entities such as spacecraft instruments and target bodies that were involved in the collection of the data.

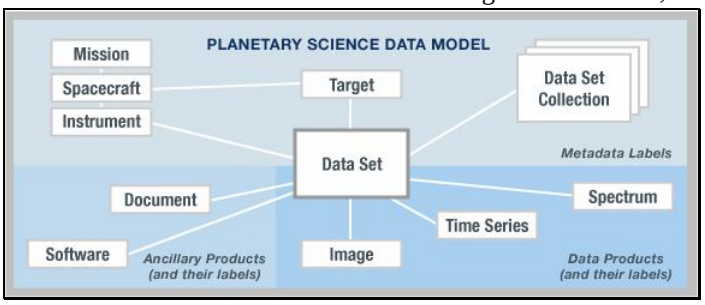

Figure 1 – The PDS Data Model

The development of the data model occurred over a period of about three years and included extensive interviews with planetary science domain experts by data management professionals. The data model was initially captured using a data dictionary and hierarchical structure diagrams, focusing on the description of planetary science entities,

their attributes and relationships. The model centered on data sets (i.e. collections of data products) and a data set's relationships to other planetary science entities. Figure 2 shows the progression of the data model's development, from structure diagrams, through the Entity-Relationship model, and then implementation in a relational schema. Finally, in order to distribute the catalog information on archive volumes as text files, the Object Description Language (ODL) was developed.

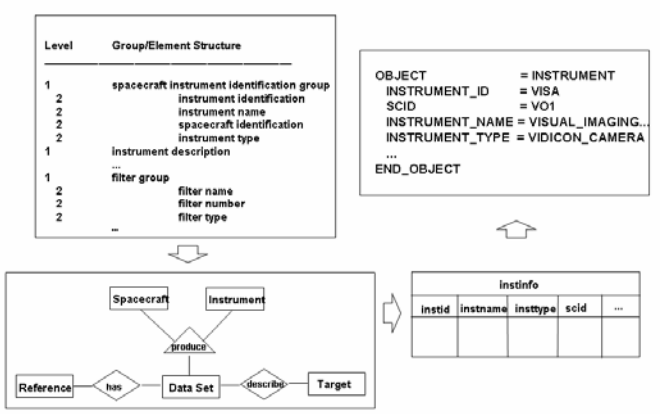

Figure 2 - Data Model Development

2.1 Ontology Development

An ontology is the product of an attempt to formulate an exhaustive and rigorous conceptual schema about a domain. It is typically a hierarchical data structure containing all the relevant entities and their relationships and rules within that domain (e.g., a domain ontology). [4].

The development of the PDS ontology was relatively easy since the PDS catalog and schema contained the essential elements, the planetary science domain object classes, their attributes, and relationships. Stanford's ontology tool, Protégé [11] was used to capture the object classes and their attributes from both the relational schema and the PDS data dictionary. Object relationships were then captured by analyzing foreign key constraints and SQL joins written for catalog applications. Since some modeling information is typically lost when implementing a relational schema, the initial interview and structure chart documentation was also used to refine the ontology. Figure 3 shows a portion of the resulting ontology as displayed by the Protégé tool, focusing on the Data Set class, its attributes (slots), and relationships. Useful modeling information often not

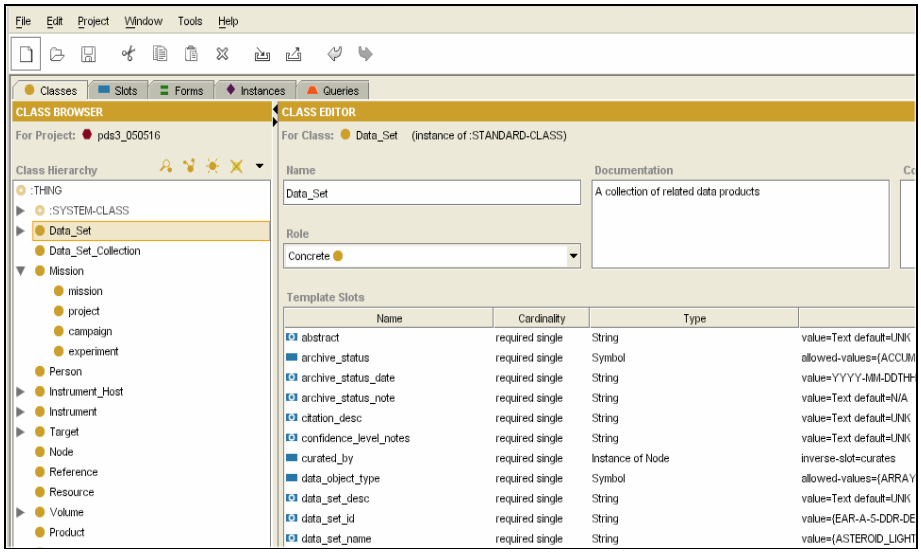

Figure 3 - Data Set Class

available in traditional modeling tools, such as subclass relationships, relationship cardinality, and inverse relations are displayed. To be consistent with Semantic Web trends the ontology was modified to include information architecture concepts from the Object-Oriented Data Technology (OODT) project. [1, 2, 3] These include broad-scope profile attributes and their relationships that support interoperability across domains. Finally, example instances of the ontology classes were ingested into the Protégé tool to validate the ontology. It should be noted that most of the PDS Catalog data could be ingested into the Protégé tool, resulting in a PDS knowledge base. This would provide yet another alternative to the PDS catalog as a source for archive information.

2.3 Ontology Representation

The capture of the data model as an ontology has resulted in a more formal and richer specification of the planetary science domain model. It revealed both known and unknown weaknesses in the model and provided alternate methods for analyzing and documenting the model. For example the Protégé tool provides several plug-ins for producing class hierarchies and UML graphical representations. A UML diagram resulting from an XML Metadata Interchange (XMI) [10] export is illustrated in Figure 4. Also since essentially all aspects of the data model have been captured in the ontology, the ontology becomes the source from which all other views of the data model can be extracted. For example, a relational schema can be generated from the ontology.

3 Semantic Web Languages

“The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." [6] In contrast, early Web development focused on people collecting information and using HTML to present the information in an organized manner for human consumption. The resulting web pages could be easily navigated by people however computers had little understanding other than how to display the information based on HTML tag semantics and how to dereference hyperlinks. For example, a computer understands that an <H2> tag should be displayed differently than an <H1> tag but does not understand that when displayed the user observes a clear hierarchical relationship between the headers.

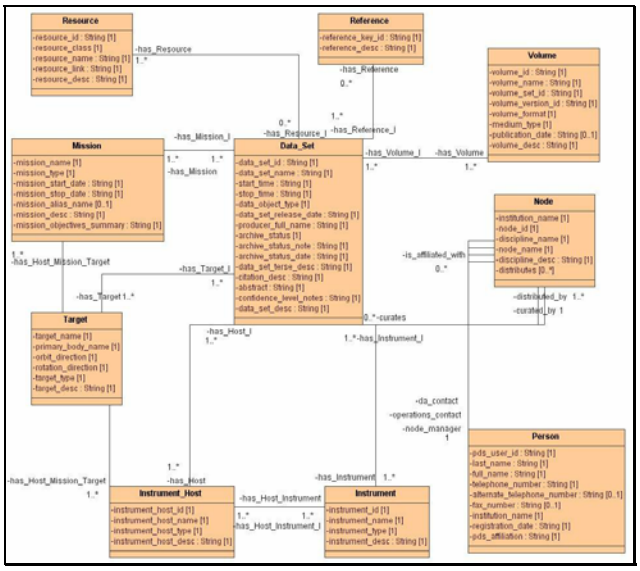

Figure 4 - UML Class Diagram

Similarly, well designed hyperlinks can illustrate vivid semantic relationships to the user while the computer is limited to understanding simple links between information items.

3.1 Resource Description Framework

The Resource Description Framework (RDF) is a language for representing information about resources in the World Wide Web. [5] RDF is based on the idea of identifying things using Web identifiers (called Uniform Resource Identifiers, or URIs), and

describing resources in terms of simple properties and property values. This enables RDF to represent simple statements about resources as a graph of nodes and arcs representing the resources, and their properties and values.

The RDF specification provides an XML-based syntax (called RDF/XML) for recording and exchanging RDF graphs that can be processed by a computer. Referring to any identifiable thing, URI's can access things that are accessible on the Web but importantly do not have to be accessible. For example, within the planetary science domain, URI's can identify image and spectrum data products that are available online as well as conceptual things such as spacecraft or instruments that are simply being described. In addition, RDF properties themselves have URIs to precisely identify the relationships that exist between the linked items. So RDF/XML provides a means to allow a machine to process ontological information about the relationship between an instrument and spacecraft.

3.2 RDF Schema

RDF properties may be thought of as attributes of resources and in this sense correspond to traditional attribute-value pairs. RDF properties also represent relationships between resources. RDF however, provides no mechanisms for describing these properties, nor does it provide any mechanisms for describing the relationships between these properties and other resources. That is the role of the RDF vocabulary description language, RDF Schema. [7] RDF Schema defines classes and properties that may be used to describe classes, properties and other resources. The RDF Schema also has an XML-based syntax, RDFS/XML.

The Protégé tool allows the export of its database content to selected semantic languages. For this prototype, class definitions, their attributes, and relationships were written to a file in RDFS/XML format, a portion of which is illustrated in Figure 5. The PDS ontology had been refined for this effort to focus on the subset of classes represented in the PDS Catalog interface. As shown in the figure, the Data_set class is defined as a subClassOf Resource and it has the properties archive_status and data_set_name. It should also be noted that for this prototype, much of the relationship information is not expressed in the RDF schema, even though it was modeled in the ontology. The goal of this prototype was to demonstrate simple facet-based search which is accomplished by using relational foreign keys. Future work will include and use the relationships modeled in the ontology.

```
<rdfs:Class rdf:about="&rdf_;Data_set"
    rdfs:label="Data_set">
  <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>

<rdf:Property rdf:about="&rdf_;archive_status"
    rdfs:label="archive_status">
  <rdfs:domain rdf:resource="&rdf_;Data_set"/>
  <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>

<rdf:Property rdf:about="&rdf_;data_set_name"
    rdfs:label="data_set_name">
  <rdfs:domain rdf:resource="&rdf_;Data_set"/>
  <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
```

Figure 5 - RDFS/ XML for a PDS Data Set

4 Semantic Search

Several academic research efforts use RDF/RDFS knowledge bases to provide semantic search capabilities. The SIMILE Project deals with applying semantic web technologies to digital libraries and providing the capability to browse and search arbitrary RDF datasets. It also supports different user interface scenarios useful to end-users, digital librarians, and metadata analysts. [8] The Simile/Longwell suite includes web-based RDF browsers that allows the user to browse and search arbitrarily complex RDF datasets using different styles including an end-user friendly view (where all the complexity of RDF is hidden) an RDF-aware view (where all the details are shown). For this prototype, the Simile/Longwell suite was chosen to provide facet-based search. Lucene is included in the suite provide text-based search. [9]

```
<rdf_:Data_set rdf:about="&rdf_;vo1/vo2-m-vis-2-edr-v2.0"
  rdf_:data_set_name="VO1/VO2 MARS VISUAL IMAGING …"
  dc:title="VO1/VO2 MARS VISUAL IMAGING SS …">

  <rdf_:target_name>
    <rdf:Description rdf:about="&terms;mars">
      <rdfs:label>MARS</rdfs:label>
    </rdf:Description>
  </rdf_:target_name>

  <rdf_:archive_status>
    <rdf:Description rdf:about="&terms;archived">
      <rdfs:label>ARCHIVED</rdfs:label>
    </rdf:Description>
  </rdf_:archive_status>
</rdf_:Data_set>
```

Figure 6 – RDF/XML for a Viking Image Data

As previously mentioned, the PDS ontology was exported from Protégé to a RDFS/XML data file. This represents the "schema" for the application and provides class and relationship information. The data for the application is contained in the RDF/XML file. For this application a Java program was written to extract for format the data set, instrument, and other entity information from the PDS Catalog database. Figure 6 illustrates the a portion the RDF/XML describing a Viking image data set and shows the data set name, the target body, and the status of the data set. Again notice that the relationship between the data set and target classes is represented using a notation that captures relational foreign keys.

The build of the Longwell semantic search application is accomplished by specifying the RDFS/XML file as an ontology, the RDF/XML file as data, and the object attributes and values to be used as "facets" in a set of configurations files. The build process produces a .war file for deployment as a web application. Figure 7 illustrates the resulting user interface where users restrict searches using an arbitrary combination of text input and facet selections. For example, the three data sets displayed are the result of two restrictions, archive_status=ARCHIVED and target_name=TITAN. The query results are displayed in the manner specified in the application build configuration files. The listed attributes are a subset of the attributes available from the RDF/XML data file and must include the attributes that have been identified as facets. The user can request the display of the entire RDF definition by clicking to the Knowle RDF navigator via the blue triangle.

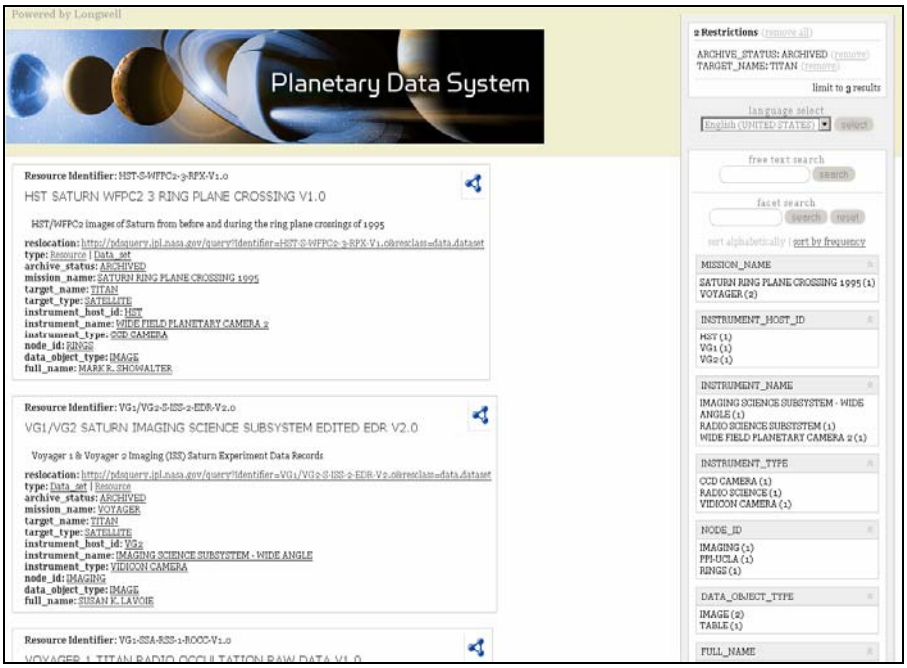

Figure 7– The Semantic PDS Search Interface

Another important restriction is “type”, or the ontological class. Since the results in the figure are only data sets, the “type” facet has been removed from the restrictions menu. For the prototype, the source RDF data file included all 1066 data sets and associated entities including thousands of targets. The search operation typically completes in two or three seconds, however large result sets will take longer to download.

5 Future Work

The Semantic PDS prototype is currently being reviewed as an alternative search interface for PDS Catalog information. Augmentations being considered include adding all attribute definitions from the Planetary Science Data Dictionary. This would more fully define the namespace specified in the PDS ontology and provide the means for users to click to the definitions of attributes that are listed in the search results. This would also provide an alternative interface for the data dictionary.

Each data set in the PDS archive represents one or more data product types, each with multiple instances. For example the Mars Viking Image data set has one image product type and about 49,000 images. To provide product level search for all products in the PDS archive using traditional database development methodologies, it would require at least a thousand product catalogs since each product type requires a unique database schema. The use of RDF/XML and RDFS/XML formatted data files eliminates the need for multiple schemas since multiple instances of these files can be loaded into a single triple-store engine. Some preliminary work is being done to explore this approach.

To support ongoing PDS data engineering, an engineering knowledge base could be developed using the concepts from the data product search described above and include example instances of PDS metadata from the archive. For example, a data provider designing a label for a new instrument product will often want to review labels from prior products for similar instruments or find all labels where a particular attribute was used.

6 Conclusion

The Planetary Data System archives data for the planetary science community. Although the total data volume in the archive is not large relative to other science domains such as Earth Science, the planetary science domain is very complex, involving dynamic contexts within which the data is collected - orbiting target bodies, moving instrument platforms, and many frames of reference. The early development of the PDS data model enabled the creation of a data archive consistent in its structure, meaning, and organization as well as rich in descriptive information.

The advent of semantic web technologies provides a means for capitalizing on this knowledge base and thereby making the planetary science archive available to a wider range of customers in increasingly more intuitive and sophisticated ways. Semantic Web technologies also suggest the means to support correlative science across science disciplines, missions, and instruments since they were designed to support inter-operability among digital assets. The Simile/Longwell suite for example allows a single knowledge base to be built using multiple and diverse ontologies and data sets. Large scale data system interoperability can be now be envisioned where “semantically aware” software agents reason about and process distributed science data repositories.

The Semantic PDS prototype demonstrates the ability to quickly develop facet- and text-based searches by leveraging existing domain catalogs. Even though designed for and implemented using relational database technology, the resulting prototype demonstrates quick development, easy deployment, and functionality surpassing that available in traditional form-based database interfaces. The prototype also suggests the potential use of semantically aware software agents to assist scientists in gleaning existing space science archives.

References

1. Crichton D, Hughes JS, Hyon J, Kelly S., “Science Search and Retreval using XML”. Proceedings of the 2nd National Conference on Scientific and Technical Data, National Academy of Science, Washington D.C, 2000.
2. Kelly S, Crichton D, Hughes J.S., “Deploying Object Oriented Data Technology to the Planetary Data System”, Proceedings of the 34th Lunar and Planetary Science Conference 1607, 2003.

3. Consultative Committee on Space Data Systems, “Space Information Architecture”, White Paper, Information Architecture Working Group. February 2004, in press.
4. Wikipedia, The Free Encyclopedia, “Ontology (computer science)”, http://en.wikipedia.org/wiki/, August 2005.
5. RDF Primer, W3C Recommendation, http://www.w3.org/TR/rdf-primer/, 10 February 2004.
6. The Semantic Web, Scientific American, Tim Berners-Lee, James Hendler, Ora Lassila, May 2001.
7. RDF Vocabulary Description Language 1.0: RDF Schema W3C Recommendation, 10 February 2004.
8. The SIMILE Project, http://simile.mit.edu/longwell/index.html.
9. Apache Lucene, The Apache Software Foundation, http://lucene.apache.org/java/docs/index.html.
10. XML Metadata Interchange (XMI), Object Management Group, http://www.omg.org/technology/documents/formal/xmi.htm.
11. Protege, Stanford Medical Informatics, http://protege.stanford.edu/.